Analysis of Causality Metrics for Determining Gene Regulatory Networks

A bachelor's thesis report
submitted in partial fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY
in
CHEMICAL ENGINEERING

Submitted by
Parth Iramani
160107030

Under the Supervision of
Dr. Resmi Suresh
November 2019

Department of Chemical Engineering

Indian Institute of Technology Guwahati

DECLARATION

I certify that

1) This work submitted by me is authentic and original, and this study has been conducted by me under the guidance of my supervisor.
2) This work has not been submitted to any other Institution for completion of a degree/diploma.
3) I have followed the Institute guidelines in documenting the work and preparing the report.
4) I have conformed to the guidelines and rules provided in the Institute's Ethical Code of Conduct.
5) I have given due credit to the authors whose research has been pivotal to this work by citing them in the report and providing their details in the reference.

Parth Iramani
Roll No. 160107030
Bachelor of Technology
Department of Chemical Engineering
Indian Institute of Technology, Guwahati
Guwahati, Assam, 781 039

CERTIFICATE

This is to certify that the work submitted in the thesis titled “Analysis of Causality Metrics for Determining Gene Regulatory Networks” by Parth Iramani (160107030) has been carried out under my supervision and guidance, and this work has not been submitted elsewhere for award of a diploma/degree.

Dr. Resmi Suresh
Assistant Professor,
Department of Chemical Engineering
Indian Institute of Technology, Guwahati
Guwahati, Assam, 781 039

ACKNOWLEDGEMENT

I would like to thank Dr. Resmi Suresh (Department of Chemical Engineering, IIT Guwahati, for her consistent and resolute support and guidance throughout the project. She has always motivated me to push my boundaries and give the best results possible. The staff of the Chemical Engineering Department, as well as my peers, who have helped me in this project, have my sincerest gratitude.

Parth Iramani
Roll No. 160107030
Bachelor of Technology
Department of Chemical Engineering
Indian Institute of Technology, Guwahati
Guwahati, Assam, 781 039

Contents

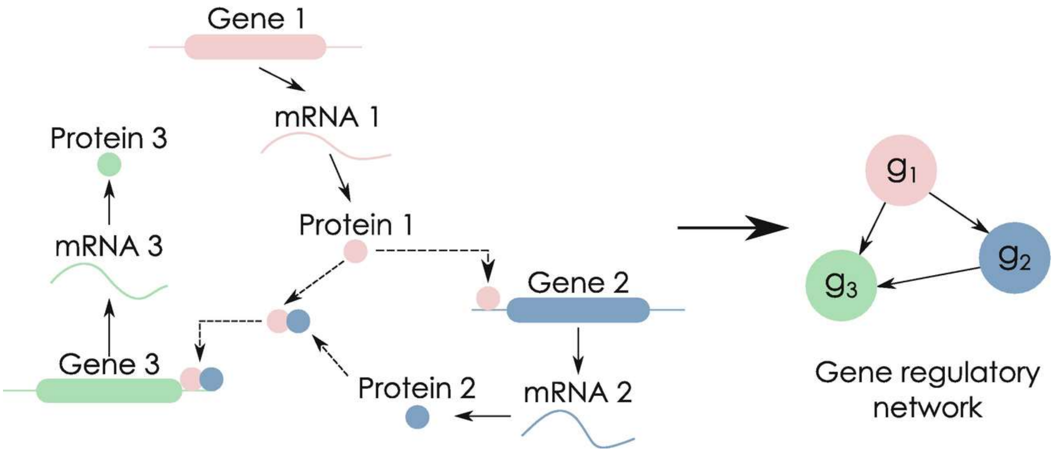
Abstract

Gene Regulatory Networks play a pivotal role in many cellular pathways and processes. Gene Expression Data can be used to determine regulatory relations among genes, greatly enhancing our understanding of cellular operations and functions. Inference of Gene Regulatory Networks is crucial for solving many biomedical and biological issues. Causality algorithms are a potential method for determining Gene Regulatory Networks from gene expression data. In this study, we analyze three different types of causality metrics – Cross-Correlation, Granger Causality, and Transfer Entropy – and their ability to predict Gene Regulatory Networks.

1. Introduction

1.1 Gene Regulatory Networks

A Gene Regulatory Network (GRN) is a set of molecular regulators, which interact with each other as well as other substances within a cell to regulate mRNA and protein gene expression levels. These regulators mostly comprise RNA, DNA, and protein complexes. The regulators can interact with each other, can activate or inhibit other regulators or complexes, and can even regulate each other indirectly. An example of a Gene Regulatory Network can be seen in Fig 1.1.



(Fig 1.1 Example of Gene Regulatory Network)
Image Credit: Gene Regulatory Network Inference: An Introductory Survey

In practice, Gene Regulatory Networks are inferred from gene expression data. (Butte et al., 2000; Liang et al., 1998; Friedman, 2004; Zhang et al., 2011) have developed different methods and algorithms to determine GRNs from gene expression data. Many of these methods have already been used to better understand diseases.

Gene Regulatory Networks have gained tremendous popularity, with much research interest being directed towards building GRN inference algorithms. GRNs are crucial in understanding many cellular operations and functions, and they can also facilitate our fundamental comprehension of disease initiation and progression. For example, GRNs are vital in understanding the disease mechanism across different types of multiple sclerosis classes (Kaznessis YN, 2007).

1.2 Causality and Causal Models

Causality refers to the cause-and-effect relations between two objects or variables. A causal model (or Causal Network) is a conceptual model that describes the causal relations between the control, dependent, and variables within a system. A causal model is often depicted as a directed graph; the nodes represent the systems' variables, and the directed edges signify a cause-effect relationship between the objects or variables. This representation of a causal model is called a causal map.

To predict a causal relation between two objects or variables, we need a causality metric. In recent works, metrics such as Shannon entropy and mutual information have been used. Other popular metrics include Granger Causality and Cross-Correlation. Transfer Entropy as a causality measure has been rising in popularity in recent studies. Using transfer entropy as a causal metric (Suresh et al., 2018) provide a mathematical framework for generating causal networks.

In this study, we analyze and compare three causal metrics – Cross-Correlation, Granger Causality, and Transfer Entropy – in terms of the accuracy of generating GRNs from gene expression data.

2. Literature Review

2.1 A hierarchical approach for causal modeling of process systems

Cause-and-effect inference is crucial for failure detection and hazard and risk assessment, requiring improvement and utilization of causal models for automated approaches. A critical disadvantage of data-driven models for fault analysis is their inability to explain their predictions and decisions. Causal models have a definite advantage here because of their inherent cause-effect reasoning. The authors have presented a hierarchical approach to building causal models using transfer entropy as a causal measure. Also suggested is an algorithm that abates the results of indirect causal interactions using reachability and adjacency matrix.

2.2 Integration of Steady-State and Temporal Gene Expression Data for the Inference of Gene Regulatory Networks

The authors develop cMIKANA, a regression algorithm for inference of Gene Regulatory Networks from gene expression data. They show that time-series and steady-state data can be successfully used together to determine GRNs using cMIKANA. They also provide a method for simulation of gene expression data of human umbilical vein endothelial cells (HUVECs), which are then used to predict GRNs using cMIKANA.

2.3 Investigating Causal Relations by Econometric Models and Cross-Spectral Methods

The authors have introduced mathematical definitions for feedback and causality, and use two-variable simple systems to test their definitions. The authors suggest that the problem of apparent immediate causality is caused either by the lag in recording information, or that all the possible causal variables are not included in the dataset. They have provided a generalization of the result that the cross-spectrum between two results can be bifurcated into parts relating to single causal arms of a feedback state.

3. Theory

3.1 Gene Expression Data

Gene expression describes a process in which gene information is used for the synthesis of a functional gene product. These products are usually proteins or protein complexes, but in non-protein-coding genes, they can be functional RNA.

Conceptually, gene expression data (or DNA microarray data) alludes to a gene-sample matrix. Gene expression data can either be steady-state or time-series. It is conventionally stored as an array, or a time-series data object in case of time-series gene expression data, wherein a column corresponds to the recorded data of a single molecular regulator, and each row corresponds to either steady-state or time-stamped recordings.

3.1.1 Simulation of Gene Expression Data

For the purpose of this study, we used simulated gene expression data, for better control over dependent and independent variables. (Wang et. al., 2013) provide a method for simulating gene expression data of human umbilical vein endothelial cells (HUVECs), which mimic the microarray experiments performed by (Hurley, et. al., 2012) in generating steady-state siRNA dataset and TNF perturbation time-series dataset.

For our study, we will be working with simulated time-series gene expression data. Time-series gene expression data is simulated by sampling the changes in gene expression as a response to the disturbance in the steady-state, which in our case, is a perturbation to a single gene in the GRN.

3.2 Causality Metrics

In this study, we will be analyzing three causality metrics.

3.2.1 Cross-Correlation

Cross-Correlation measures the similarity between two signals or series as a function of the relative displacement between the two series. Cross-Correlation has many applications in electron tomography, pattern recognition, and causality. For two random vectors X and Y with probability density functions f and g, then, we can measure the difference between X and Y (or $Y - X$) by the Cross-Correlation between $f(t)$ and $g(t)$. Since Cross-Correlation can tell us the direction and magnitude of similarity between two series, it serves as an excellent causality metric.

The Cross-Correlation between two continuous functions f and g is given by:

$$(f * g)(\tau) \triangleq \int_{-\infty}^{\infty} \overline{f(t)} * g(t+\tau)dt$$

3.2.2 Granger Causality

(Granger 1969) introduced the Granger Causality test, a statistical hypothesis test for judging if a series X is useful in forecasting another series Y. Clive Granger states that it is possible to determine causality between two time-series, by measuring the ability of prior values of one series to predict the future values of the second series. A time-series X Granger-causes Y if the predictions of the future values of Y based on the past values of Y as well as the past values of X are statistically better than the predictions of future values of Y based only on the past values of Y.

Granger Causality defines causal relations based on two assumptions.

1) The cause is prior to effect
2) The cause contains unique information about the future values of effect.

With these two assumptions, to determine causal relations between X and Y we test the following hypothesis.

$$P[Y(t+1) \epsilon A \mid I(t)] \neq P[Y(t+1) \epsilon A \mid I_{-x}(t)]$$

A denotes an arbitrary non-empty set, P denotes probability, $I(t)$ denotes the entire information available at time t, and $I_{-x}(t)$ denotes the entire information available at time t excluding information about X.

If this hypothesis is accepted, then we can say that X Granger-causes Y.

3.2.3 Transfer Entropy

(Schreiber, 2000) introduced transfer entropy, an information-theoretic measure. Transfer entropy quantifies the directional transfer of information between two variables. For e.g., if we have a directed transfer of information from variable x to y, the transfer entropy is estimated as

$$\mathrm{T_{x\to y}} = \sum_i p(y_{i+h}, \boldsymbol{x}_i,\ \boldsymbol{y}_i) \ln \frac{p(y_{i+h}|\boldsymbol{x}_i, \boldsymbol{y}_i)}{p(y_{i+h}|\boldsymbol{y}_i)} \ = H(y_{i+h}|\boldsymbol{y}_i) - H(y_{i+h}\ |\boldsymbol{y}_i, \boldsymbol{x}_i)$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are embedded vectors representing $[x_i, x_{i-\tau}, \dots x_{i-(k-1)\tau}]$ and $[y_i, y_{i-\tau}, \dots y_{i-(k-1)\tau}]$, respectively. The time interval considered between the various variables in x and y is denoted by τ. The prediction horizon is represented as h, and the embedding dimensions of x and y are denoted by k and l, respectively. H denotes the conditional entropy.

This metric of information transfer is then used by the framework to quantify the information transfer between any two variables in a system. The information transfer is taken as a direct measure to estimate direct causality between the variables.

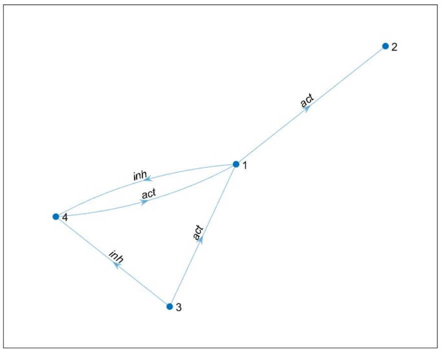
Note that for this study, we won't be using the full Transfer Entropy model provided by (Suresh et. al., 2018), and will just be using Transfer Entropy values, since we want the three metrics to be at the same level of simplicity.

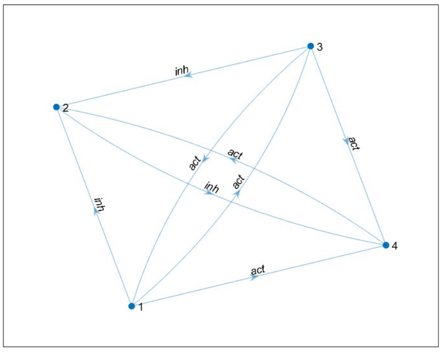
4. Process Workflow

The following workflow has been adopted to compare and analyze the three causality metrics.

4.1 Candidate Gene Regulatory Networks

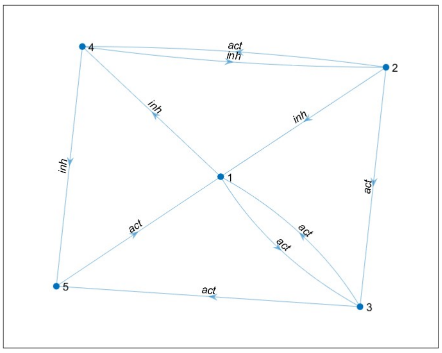
We selected eight candidate Gene Regulatory Networks, which are to be predicted. We decided the network of these GRNs beforehand. The eight candidate GRNs, named A1 to A8, are shown below. The number of molecular regulators (genes) in A1 and A2 is 4, A3 and A4 is 5, A5 and A6 is 6, A7 and A8 is 7.

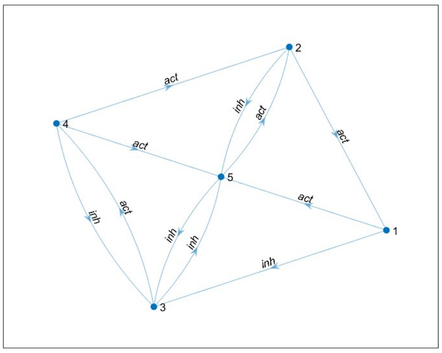




(Fig 4.1.2 Causal Map of candidate GRN A1)

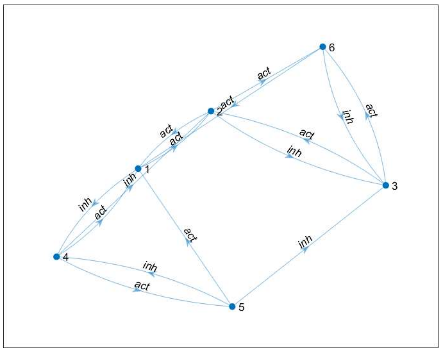
(Fig 4.1.1 Causal Map of candidate GRN A2)



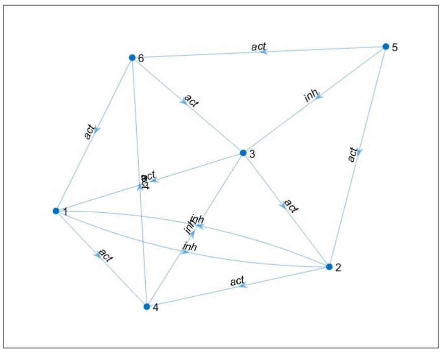


(Fig 4.1.3 Causal Map of candidate GRN A3)

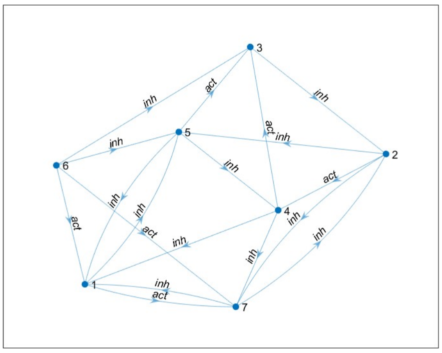
(Fig 4.1.4 Causal Map of candidate GRN A4)



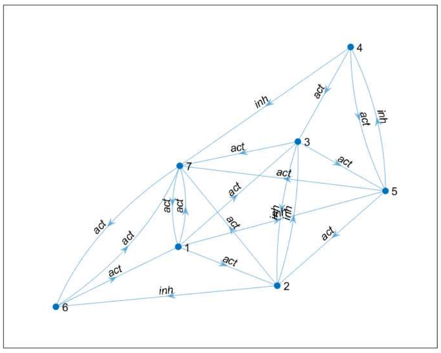
(Fig 4.1.5 Causal Map of candidate GRN A5)



(Fig 4.1.6 Causal Map of candidate GRN A6)



(Fig 4.1.7 Causal Map of candidate GRN A7)



(Fig 4.1.8 Causal Map of candidate GRN A8)

The above figures are graph representations of the GRNs, with nodes and edges. In each of these GRNs, the nodes denoted by numbers represent individual genes in the GRN. The edges represent the causal relations between the genes, with the direction being depicted by the arrow-heads, and the type of interaction between the genes (activatory or inhibitory) is shown along the edges.

4.2 Simulation of Gene Expression Data.

For each of the candidate GRNs, we simulated a gene expression dataset for a perturbation in each gene. For e.g., the candidate GRN A1 has four genes, therefore, four gene expression datasets were simulated for a perturbation in each gene. Therefore, the number of gene expression datasets generated for a candidate GRN is equal to the number of genes in the candidate FRN.

The gene expression data is simulated according to the algorithms provided by (Wang et. al., 2013) for simulating gene expression data which mimic the microarray experiments performed by (Hurley, et. al., 2012) in generating steady-state siRNA dataset and TNF perturbation time-series dataset.

4.3 Determination of GRNs using simulated gene expression datasets.

After generating gene expression datasets, we use the three causality metrics to determine GRNs for the respective datasets.

For each simulated gene expression dataset, we generate GRNs using each of the three causality metrics – Cross-Correlation, Granger Causality, and Transfer Entropy. Therefore, for each simulated gene expression dataset, we have three GRN predictions, corresponding to the three causality metrics. Since these GRNs are determined from the gene expression data simulated from the perturbation of a single gene in the candidate GRN, we named these GRNs as single-gene-perturbation GRNs.

For each of the three causality metrics, we adopted the same approach. So, if we are to consider Cross-Correlation; for a gene expression dataset, we took all possible pairs of genes representing the relation between the genes in the pair (except the pairs which represent the relation of a gene upon itself) and found the causal relation between these pairs using Cross-Correlation. These causal relations between pairs of genes were then combined together to form a GRN. The same holds true for Granger Causality and Transfer Entropy.

4.4 Final Predictions for candidate GRNs

Finally, all the single-gene-perturbed GRNs corresponding to their respective candidate GRNs determined using Cross-Correlation were combined to form the final prediction for the candidate GRNs. The same was done for Granger Causality and Transfer Entropy. Two methods of combining the single-gene-perturbation GRNs gave the best results.

4.4.1 Method 1 (Average of causal relation values)

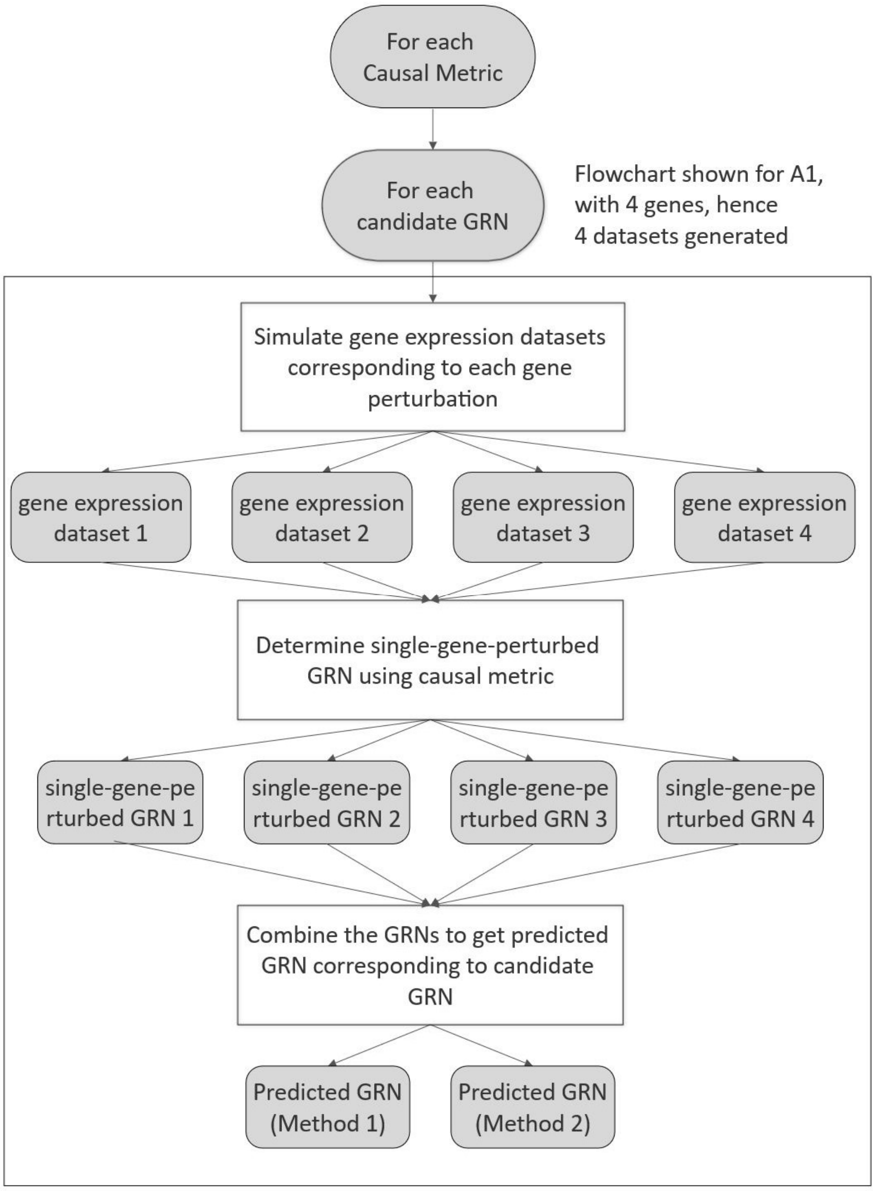
In the first method, we combined the single-gene-perturbation GRNs corresponding to the candidate GRNs by averaging out the causal relation values corresponding to each gene-pair relation found by the causal metric.

4.4.2 Method 2 (Absolute maximum/minimum of causal relation values)

In the second method, we combined the single-gene-perturbation GRNs corresponding to the candidate GRNs by using the absolute maximum (minimum in the case of Granger Causality) values of the causal relation values corresponding to each gene-pair relation found by the causal metric.

4.5 Flow Diagram

To following flow diagram (Fig 4.5) to better elucidate the workflow.



For each Causal Metric
For each candidate GRN
Flowchart shown for A1, with 4 genes, hence 4 datasets generated
Simulate gene expression datasets corresponding to each gene perturbation
gene expression dataset 1
gene expression dataset 2
gene expression dataset 3
gene expression dataset 4
Determine single-gene-perturbed GRN using causal metric
single-gene-pe rturbed GRN 1
single-gene-pe rturbed GRN 2
single-gene-pe rturbed GRN 3
single-gene-pe rturbed GRN 4
Combine the GRNs to get predicted GRN corresponding to candidate GRN
Predicted GRN (Method 1)
Predicted GRN (Method 2)

(Fig 4.5 Workflow flowchart)

5. Results

5.1 Evaluation of Causal Metrics

The causal metrics have been evaluated on the basis of how close the final predictions of the candidate GRNs are to the actual candidate GRNs. The evaluation metric used is accuracy, therefore the evaluation of the causal metric is based on the percentage of causal relations in the predicted GRN that match the candidate GRN.

5.2 Result Table

The final results have been tabulated in the below table (Table 5.2)

GRN	**Cross-Correlation**		**Granger Causality**		**Transfer Entropy**	
	Method 1	*Method 2*	*Method 1*	*Method 2*	*Method 1*	*Method 2*
A1	50.00%	58.33%	41.67%	33.33%	50%	41.67%
A2	41.67%	41.67%	41.67%	58.33%	41.67%	25.00%
A3	40.00%	30.00%	50.00%	20.00%	35.00%	20.00%
A4	45.00%	45.00%	50.00%	40.00%	45.00%	35.00%
A5	36.67%	36.67%	40.00%	36.37%	33.33%	30.00%
A6	53.33%	40.00%	43.33%	40.00%	30.00%	40.00%
A7	52.38%	38.10%	57.14%	52.38%	47.62%	33.33%
A8	47.62%	42.86%	45.24%	28.57%	45.24%	35.71%
Overall Accuracy	45.83%	41.58%	46.13%	38.66%	40.98%	32.59%

(Table 5.2 Accuracy Scores)

6. Conclusion and Future Work

We analyzed the ability of three causal metrics – Cross-Correlation, Granger Causality, and Transfer Entropy – to determine Gene Regulatory Networks from simulated gene expression data. Comparing the accuracy results for the three causal metrics, we notice that no metric performs consistently better than the other two metrics. Comparing overall accuracy results, Granger Causality and Cross-Correlation perform similarly, while Transfer Entropy lags behind. This is to be expected, as we took only the Transfer Entropy values for determining the causal relation values, make the algorithm simpler as to be on the same level as Cross-Correlation and Granger Causality.

The accuracy scores of the predicted GRNs using the three metrics are acceptable, yet far from being exceptional. In future work, we aim to improve the accuracy of these algorithms. We need to further analyze and understand the reason for the deviation from the expected candidate Gene Regulatory Networks. We also plan on extending this study by observing more complex GRNs. Since in this study we have restricted the Transfer Entropy model, future studies can also be conducted to analyze the results of the full Transfer Entropy model.

References

[1] Butte, A., Tamayo, P., Slonim, D., Golub, T., and Kohane, I. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc. Natl. Acad. Sci. U.S.A. 97, 12182–12186. doi: 10.1073/pnas.220392197

[2] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. Science 303, 799–805. doi: 10.1126/science.1094068

[3] Granger, C. W. J. “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods.” Econometrica, vol. 37, no. 3, 1969, pp. 424–438. JSTOR.

[4] Hurley D, Araki H, Tamada Y, Dunmore B, Sanders D, et al. (2012) Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. Nucleic Acids Research 40: 2377–2398

[5] Kaznessis YN (November 2007). "Models for synthetic biology". BMC Systems Biology. 1: 47. doi:10.1186/1752-0509-1-47.

[6] Liang, S., Fuhrman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac. Symp. Biocomput. 1998, 18–29

[7] M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs and N. F. Thornhill, "Finding the Direction of Disturbance Propagation in a Chemical Process Using Transfer Entropy," in IEEE Transactions on Control Systems Technology, vol. 15, no. 1, pp. 12-21, Jan. 2007.

[8] Resmi Suresh, Abhishek Sivaram, Venkat Venkatasubramanian, A hierarchical approach for causal modeling of process systems, Computers & Chemical Engineering, Volume 123, 2019, Pages 170-183, ISSN 0098-1354.

[9] Schreiber, T., Schmitz, A., 20 0 0. Surrogate time series. Physica D 142 (3-4), 346–382.

[10] Wang YK, Hurley DG, Schnell S, Print CG, Crampin EJ. Integration of steady-state and temporal gene expression data for the inference of gene regulatory networks. PLoS One. 2013.